SCIENTIFIC REPORTS
nature research

OPEN

Using already-solved cases of a mass disaster event for prioritizing the search among remaining victims: a Bayesian approach

Inés Caridi[1*], Enrique E. Alvarez[2], Carlos Somigliana[3] & Mercedes Salado Puerto[3]

This work presents a new method for assisting in the identification process of missing persons in several contexts, such as enforced disappearances. We apply a Bayesian technique to incorporate non-genetic variables in the construction of prior information. In that way, we can learn from the already-solved cases of a particular mass event of death, and use that information to guide the search among remaining victims. This paper describes a particular application to the proposed method to the identification of human remains of the so-called *disappeared* during the last dictatorship in Argentina, which lasted from 1976 until 1983. Potential applications of the techniques presented hereby, however, are much wider. The central idea of our work is to take advantage of the already-solved cases within a certain event to use the gathered knowledge to assist in the investigation process, enabling the construction of prioritized rankings of victims that could correspond to each certain unidentified human remains.

The process of identification that guides searches in contexts such as disaster victim identification (DVI), missing person identification (MPI), migration and other situations of violence (OSV) requires the collection of background information from different sources (e.g. legal courts documents, testimonies from survivors, witnesses and families of the missing)[1]. The identification process is essential not only for the sake of Justice and for humanitarian reasons[2] but also to offer answers to victims' families and friends[3–6]. The process of identification usually includes both, (*i*) the construction of hypotheses of identity from the analysis of such background information that needs to be evaluated at a later stage through genetic evidence, and (ii) the validation of the information gathered from a genetic DNA-led process through the comparison of the ante-mortem and post-mortem information. It is our aim in this paper to describe a general method which could contribute to the investigation process by taking advantage of the already-solved cases of a particular mass death event, to use that elicited knowledge for guiding new searches of related unidentified human remains (UHR). Whenever a pattern does exist within the already-solved cases, the method presented here allows us to make predictions in the identification process of the cases still unsolved, and it also makes it possible to minimize any bias from the researcher. Predictions are understood as the act of prioritizing some individuals over others to be more likely related to certain UHR within the same event. The available information is: (i) information regarding the context of the mass death event, such as date and place in which the event has occurred and the total number of victims, (ii) a database with information of reported victims who are potential candidates to correspond with a set of UHR. That database also includes non-genetic variables amenable to be modeled mathematically in the search for patterns, and (iii) information of the set of already-solved cases within the same mass event (i.e. cases already identified). Thus, in essence, we have two lists: the set of potential victims whose human remains have not been found yet (which we denote in the sequel List 1) and the set of UHR which have not yet been identified (which we denote by List 2 in the sequel). The main idea of this work is to update an initial instance of knowledge, in which all possible victims are equally likely to correspond with certain UHR within the same mass death event, into a new instance of knowledge, in which some victims are more likely to match certain UHR, in the light of the data resulting from the already-solved cases. We accomplish that updating process mathematically with the aid of Bayesian techniques. Those are used in

[1]Instituto del Cálculo and CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [2]Facultad de Ingeniería, Universidad Nacional de La Plata and CONICET, Buenos Aires, Argentina. [3]Equipo Argentino de Antropologia Forense, Buenos Aires, Argentina. *email: ines@df.uba.ar

two ways: i) to elicit background information about the probabilities of matches from UHRs to potential victims in the form of a prior distribution, and ii) to update that prior into a posterior distribution. In this work, we opt to evaluate the results of our method by using cross-validation techniques, for which we introduce two measures of goodness of fit, that we call *Discriminating Power* (*DP*), and *Efficacy Rate* (*E*), as defined in Section *Validation techniques* below. Basically, the *DP* measures the ability of a model fitted only from a *training sample* to update the probabilities of matches using a *validation sample* (e.g., fresh cases). In turn, the *E* measures how *informative* the model, based on the training sample, results for the validation sample. As outputs, the model firstly identifies the best set of informative non-genetic variables to learn from the already-solved cases of a mass event and provides the optimal way of partitioning those variables. Each of those partitions entails a candidate model that will be evaluated through cross-validation. Secondly, the model gives an analytical expression that, when applied to the best partition of the set of non-genetic variables, which implies a partition of the victims (List 1) in subsets, generates probabilistic scores of victims for unidentified human remains of related cases. It is noteworthy that scores are a function of the data and then, each new piece of information (from both the victims database or the set of already-solved cases) could generate new results.

In the statistical analysis of genetic matched data, the probability of the event "the DNA sample is related to the victim" is usually compared to the alternative "the DNA sample is not related to the victim" by calculating odds ratios. The Bayesian approach updates the quotient of this prior odds (built on non-genetic evidence) to obtain the posterior quotient (posterior odds) by multiplying the prior odds by the likelihood ratio (built on genetic evidence)[7–10]. Very low prior odds and/or low quality and quantity of DNA, both of the material extracted from bones and DNA samples from victims' relatives are some of the causes which could hinder reaching the required threshold of identification that could be reached by improving biological reference samples. For this reason, it is important to prioritize the possible victims of the event in order to allocate efforts to obtain certain DNA samples from close family donors. A good guide on how to prioritize the research of unsolved-cases would, in principle, allow substantial savings in time and resources. As an example of the use of non-genetic information to prioritize or highlight the searches, the collaborative online platform called "Reuniting Families" uses non-genetic information to flag samples of interest that are manually examined by experts using other available data[11].

During the last Argentine dictatorship, several circuits of Illegal Detention Centers (IDC) were set up in different locations throughout the country. There, thousands of persons were illegally held without any sort of legal guarantees, tortured and most of them killed. Their unidentified bodies were buried in individual or common graves either within official cemeteries or in clandestine mass graves at military or police compounds. Even today, the fate of most of those disappeared people's remains is still unknown[1]. Missing people have come to be known as "The Disappeared". Since 1984 the Argentine Forensic Anthropology Team (EAAF) has been working on the identification of the disappeared using a multidisciplinary approach[12–14]. The identification process related to these events which occurred over 35 years ago presents important challenges. First, the information from reports is incomplete; second, even though there are reference samples available from only approximately half of the victims' families, in some cases the samples belong to distant relatives with a weaker DNA connection. This is because in many cases, relatives who are very informative from a genetic point of view (e.g., mothers and fathers) have already died or are now very old. Moreover, in some cases, several members of the same family were disappeared, for example, the "missing grandchildren"[5].

The need to generate hypotheses of identity for the recovered human remains and the fact that in some mass events there are sets of already-solved cases, triggered an interest in developing a model that mathematically systematizes information obtained by already-solved identifications, in such a way that it could be used in the search and generation of new hypotheses of identity for related unsolved cases. A Bayesian model is proposed to learn from already-identified cases and to generate a probabilistic ranking of victims for unidentified related cases. This ranking is, by nature, dynamic, as it can be sequentially updated every time new information from sets is appended (List 1 of victims, List 2 of unidentified remains and the set of solved-cases). In that way, the model allows the identification of the most informative subset of non-genetic variables to detect patterns, which beside achieves results that are significantly better than those obtained just by chance. In other words, once the updating has been formalized, a ranking of suitable victims for the recovered skeletal remains is produced. An important advantage of this method is that it minimizes any bias in the investigation of related cases.

Results

The methodology presented here was applied to events within the context of the last dictatorship in Argentina, such as the so-called Massacre of Fátima. On August 20th, 1976, ten women and twenty men were killed in the township of Fátima, Province of Buenos Aires, Argentina. Hence, it was a well-delimited event, both geographically and temporally, and it involved a well-known number of people. Until now, 24 out of 30 of them have been identified at different stages. Within all the victims of the dictatorship in Argentina, it is possible to select a subset List 1 of individuals to form the set of candidate victims that could correspond to the unidentified remains List 2 associated with the particular event, such as Fátima.

The first challenge to build scores within the set of eligible victims is to select the non-genetic informative variable or variables to learn from the already-solved cases. With this in mind, a partition within the complete range of non-genetic variables was defined on a grid, which entails a grouping of List 1 into some subsets. Every cell on the grid is associated with a combination of values of those variables.

For simplicity, let us assume for a moment that there were only two variables, of geographical and temporal nature, respectively, such as those related to the place and date of kidnapping of the victims. In that case, boxes on a grid represent GeoTemporal cells in the sequel. For data analysis, narrowing or enlarging the time window widths entails parameters to be calibrated later seeking a balance for an optimal division based on a combination of *DP* and *E*. Thus, every identified individual and every victim are placed in a GeoTemporal cell (subset) based on the date and place of their kidnapping. Then, the main idea is to update the importance of each GeoTemporal

cell (GeoTemporal probabilities), which represents the probability that the unidentified human remains of the same event may correspond to a victim belonging to this cell. In the initial instance of knowledge of the problem, GeoTemporal probabilities have to be consistent with what is known before the identifications of the event were made, that is, all the possible victims are equally probable, and therefore those cells with more victims will be more likely. This fact means that the prior probabilities of the cells (before having data from identifications) are proportional to the number of possible victims from every GeoTemporal cell. Moreover, within a given cell, we consider that there are a-priori equal probabilities for each of the individuals belonging to this cell. Updated probabilities of GeoTemporal cells will result from a combination of the prior probabilities and the information of the set of GeoTemporal cells observed from the identified cases of that particular event. Then, for certain unidentified human remains from the same event, a probability score to every possible victim is assigned.

Thus, the first step to formulate the victims rankings scores consists of the construction of an informative probability distribution among possible victims learning about the already-solved cases of the same event, based on non-genetic variables. The second step consists of using those probabilities to prioritize some victims as more likely to correspond with UHR. The last step, left for evaluation by forensic experts, is to physically evaluate the rankings to assist the investigation and to build new hypotheses. The information they provide after this process could be incorporated in a new sequential step of prior elicitation. A tool to implement this feedback from forensic researchers is under construction and will be available in a free and open interface[15].

Bayesian framework. In this Section, we introduce the notation to be used in the rest of the paper. Let S represent recovered skeletal or human remains of an individual, although still unidentified (unidentified human remains UHR) associated to a particular mass event of death (this is an element of what we have called List 2 in the *Introduction*). We also denote the set of possible victims that corresponds with S by $V = \{v_1, v_2, \ldots, v_N\}$ (this is what we call List 1 in the Introduction), and denote $\#V = N$ (being # the total number of elements of the set). Then $P(v_i \text{ is } S)$ denotes the probability that the i-th possible victim (v_i) corresponds to UHR S. Hence the expression $P(v_i \text{ is } S) = 0$ means that with certainty v_i is not the UHR S, while $P(v_i \text{ is } S) = 1$ means that with certainty, v_i corresponds with S.

As mentioned, the purpose of this paper is to update the probability $P(v_i \text{ is } S)$, as computed in an initial instance of knowledge, which ignores any information from the observations of the already-solved cases, into some updated probabilities using information from the already-solved cases. The initial and updated probabilities are called *prior* and *posterior* probabilities, respectively, in Bayesian Statistics. They are denoted mathematically by $P(v_i \text{ is } S)$ and $P(v_i \text{ is } S|\mathscr{D})$, where the latter is a conditional probability in which the set $\mathscr{D}$ denotes the data gathered from the already-solved cases. In our presentation below, we consider that *a priori* there is ignorance concerning the event, in the sense that $P(v_i \text{ is } S) = 1/n$ is the same value for each of the i-th possible victims (a fact modelled by a discrete uniform distribution). However, after the updating process, the posterior probabilities $P(v_i \text{ is } S|\mathscr{D})$ differ and lead to a ranking of possible matches to be evaluated by forensic experts. The main idea is to use Bayesian Inference, which explains how to update some prior probabilities after appending new information, to give rise to the posterior probabilities[16].

In our case, a straightforward application of Bayes' Theorem entails

$$P(v_i \text{ is } S|\mathscr{D}) = \frac{P(v_i \text{ is } S)P(\mathscr{D}|v_i \text{ is } S)}{P(\mathscr{D})}, \tag{1}$$

Naturally, apart from the information elicited from the already-solved cases ($\mathscr{D}$), there could be another type of background information. At this stage, it is worth it to mention that Bayesian inference is now recognized as the most useful model to understand how evidence may be presented logically and impartially in legal proceedings[7], because the underlying assumptions are all explicit. In our treatment for this manuscript, those assumptions are enumerated as follows.

Assumptions of the identification problem. The assumptions considered to build probabilities for every individual of set V to correspond to UHR S of a particular mass event of deaths (or simply the *event*) are the following:

- (a) The event is well delimited geographically, temporally (the data of occurrence, called T_e, is known) and in terms of the number of killed individuals (called n_e).
- (b) There is a set of already-solved cases of the event, size n_s, which is a subset of the complete set of deaths ($n_s \leq n_e$).
- (c) It is known with certainty that recovered UHR S corresponds to one of the set of n_e individuals related to the event. Moreover, S correspond to just one of the victims, which excludes the possibility of mixed remains.
- (d) The set of possible victims that corresponds to recovered human remains related to the event under study is known.
- (e) The set of already-solved cases (n_s) is a random sample from the complete set of deaths of this mass event (n_e).
- (f) There is a known set of non-genetic variables associated with each victim (e.g., geographical, political, and time related variables).

From assumptions (a), (c) and (d) it is possible to define the set of victims that correspond to recovered human remains S related to the event, called $V = \{v_1, v_2, \ldots v_N\}$ (List 1, mentioned in the *Introduction*), as the subset of the

N_j (victims), I_j (identified)	geo1	geo2	geo3	...
temp1	35, 0	12, 2	3, 0	...
temp2	54, 9	23, 3	1, 0	...
temp3	56, 4	27, 4	6, 0	...
...	...	...	...	...

Table 1. Example values of a table of N_j (number of victims belonging to cell j) and I_j (identified cases of cell j) for some $m = 9$ GeoTemporal cells which result from using a particular temporal parameter $T = 15$ days to define the length of the cells in temporal variable, and using the division of the region in areas of interest. Three areas are shown in this example. Column geo1 and row temp1 define the cell 1, column geo2 and row temp1 define the cell 2, and so on until cell m. In this example, $N = 217$ and there are $n_s = 22$ already-identified cases. In this example, listing the cells from 1 to 9 from left to right and from top to bottom, the prior probabilities for the cells are N_j/N (0.161, 0.055, 0.014, 0.249, 0.106, 0.005, 0.258, 0.124, 0.028) and the posterior probabilities are the following values (0.013, 0.088, 0.001, 0.397, 0.134, 0.000, 0.188, 0.177, 0.002), which will be explained in the following subsection.

total set of victims who were kidnapped or missing before date T_e. In this sense, V is defined as the *effective* set of possible victims, because it does not include those victims whose probability is zero. This consideration takes into account only the assumptions enumerated above in this manuscript. In other words, it does not include either the sex or the age of the UHR S. Using assumption f), it is possible to associate each victim (both the possible victims and the identified ones in the event) with a non-genetic variable or a set of non-genetic variables, as for example the geographical area, and the period before the date of the event when the kidnapping took place. Given a variable, it is possible to define a *partition* of the set of values of this variable such as a division into blocks or cells of values of the variable. Every cell could be associated to a range of values of the variable, both for numeric or categorical variables. If two variables are considered, each cell could be associated with a pair of values; if three variables are considered, each cell could be associated with a third one, and so on. Once a partition is constructed in the form of cells, it implies a grouping of the set of possible victims in the given subsets.

GeoTemporal cells. For the sake of simplicity, the following sections will describe the statistical calculations applied to a set of variables associated with the place and date of the kidnapping of the victims. For this reason, the cells are called "*GeoTemporal*". The number of possible compatible victims for each cell is recorded. The time variable was partitioned by considering periods of length duration T, starting out from the date of the event T_e and going backward in time. The geographical variable was partitioned by considering areas of interest, associated with the relevant areas in the context of the historical phenomena under study. Then, each cell of the GeoTemporal partition will be associated with both, a geographical area and a specific range date of T days of length. To simplify notation, cells with index numbers from 1 to m are tagged. The referred cells look like Table 1.

Probabilities for every cell. Let us call $\mathscr{N}_j$ the subset of possible victims within V which has the combination of variables associated with cell j, and let N_j be the size of this set ($N_j = \#\mathscr{N}_j$). Thus, by $S \in \mathscr{N}_j$ we mean that UHR S belongs to the set $\mathscr{N}_j$ (i.e., UHR S corresponds to one of the individuals of subset $\mathscr{N}_j$). Let $P\left(S \in \mathscr{N}_j\right)$ be the probability that S correspond to an individual who belongs to cell j. In this way, $P\left(S \in \mathscr{N}_j\right)$ is the prior probability that remains S correspond to an individual who belongs to cell j, where *prior* refers to the probabilities calculated in an instance of knowledge before the data of the already-identified cases were observed; while $P\left(S \in \mathscr{N}_j \middle| \mathscr{D}\right)$ is the posterior probability that UHR S come from cell j, where posterior refers to the instance of knowledge updated after the data of the already-identified cases were observed. Posterior probabilities will be obtained from prior probabilities after learning from the experience of the dataset of identified cases of the particular mass event of death ($\mathscr{D}$). In the *Methods* section, it is described how to update mathematically prior probabilities $P\left(S \in \mathscr{N}_j\right)$ into posterior probabilities $P\left(S \in \mathscr{N}_j \middle| \mathscr{D}\right)$ in a Bayesian treatment.

Updating probabilities for every possible victim. Once the probability that remains S belong to some victim of cell j is updated (which in the sequel we denote by $P\left(S \in \mathscr{N}_j \middle| \mathscr{D}\right) =: \hat{\theta}_j^{\text{post}}$ for simplicity of notation), it is assumed that within that particular cell all the victims that are still unidentified ($\mathscr{N}_j$) have the same chance of corresponding with UHR S. For this reason, the probability that remains S correspond to the specific individual i whose non-genetic variable values are associated with cell j is:

$$P(v_i \text{ is } S|\mathscr{D}) = P_j^i = \frac{\hat{\theta}_j^{\text{post}}}{N_j} = \frac{1}{N_j}\left[\frac{\frac{N_j}{N}\left(\frac{N}{N_k} - 2\right) + I_j}{\frac{N}{N_k} - 2 + n_s}\right], \tag{2}$$

where v_i is the i-th individual from cell $j\left(v_i \in \mathscr{N}_j\right)$, and k is the cell in which there are more victims (see the section *Methods* below for details). The number of victims who are unidentified from cell j is N_j, and I_j is the total number of already-solved cases associated with cell j, such as $\sum I_j = n_s$. The value of P_j^i will be the score associated with each individual i of set V belonging to cell j to correspond with UHR S. The ranking of priorities for certain

remains S is built by listing the set of possible victims according to decreasing scores, which implies the existence of a ranking of blocks. Those victims, sharing the same score value (because they belong to the same cell), will be together in the same block.

Validation techniques. The resulting scores are dependent on the particular partition of the subspace of non-genetic variables into cells. For each subset of non-genetic variables used, different partitions could be evaluated. The set of partitions will be used to group the victims into subsets of individuals. Each of these groupings will represent a "model". Moreover, some variables, such as time, involve a way of partitioning that is in turn a function of a parameter that defines the length of the cell (parameter T) (for example, Table 1 shows example values of N_j and I_j for a particular GeoTemporal cells involving parameter $T = 15$ days). Then, we propose the implementation of a sensitivity analysis of the results for the different partitions of the set of possible victims V through cross-validation techniques, which provides information to identify the best subset of non-genetic variables to define the cells partition. It is noteworthy that the general methodology we propose in this paper is, in some sense, "*hybrid*". That is because while we opted for a fully Bayesian treatment to update the matching probabilities, we chose a classical method for model selection via cross-validation. A fully Bayesian treatment would entail placing prior probabilities for different models (in this case partitions of the non-genetic variables) in the form of the so-called Bayes factors to combine the models in the final inference. Our choice of the hybrid method is due to two reasons: (i) there is no clear way to elicit priors for different models in our context; and (ii) using cross-validation for model selection is more amenable to be included automatically in computing code, without the need of interactive analysis by a statistical practitioner. This makes the method more attractive for applications by Forensic Scientists or professionals without specific statistical training. Broad applicability outside the specific statistical community is one of the main goals we seek in the dissemination of this work.

Cross-Validation divides the data of the already-solved cases randomly into two samples: (i) a learning sample $\mathscr{D}_{\mathscr{L}}$, in which several model options are comparatively estimated, aiming at the best fit; and (ii) a validation sample $\mathscr{D}_V$, used to evaluate the models with the reserved data. The cross-validation strategy is accepted in the literature as a way to prevent overfitting (i.e., proposing a model that works very well for the learning sample but very poorly with fresh data), while providing good predictions with new data[17]. The main idea is to implement the calculations described (and detailed in *Methods*) by using $\mathscr{D}_L$ instead of $\mathscr{D}$, to generate the results and evaluate the scores of those individuals who belong to subset $\mathscr{D}_V$ to quantify the results. It is necessary to define adequate magnitudes to measure the goodness of fit. Following Hastie[18], $\mathscr{D}_L$ is selected from the original sample by taking a random statistical subset of 75 percent size of the original sample from $\mathscr{D}$ (with no replacement); and the remaining cases form subset $\mathscr{D}_V$. The key is to pretend that the reserved cases have not been identified yet and to track and observe them in the final results.

In the ranking of possible victims for a particular learning sample $\mathscr{D}_L$, it is desirable i) that those victims belonging to the validation sample ($\mathscr{D}_V$) obtain higher scores than those in the initial instance of knowledge and ii) that there are not many cases outside the validation sample that improve their scores. For those reasons, we focus on two magnitudes, i.e.:

- *Discriminating Power*, *DP*, which is defined as the fraction of the reserved cases which obtain greater scores than in the initial instance of knowledge ($\mathscr{R}_+$) with respect to the size of the reserved sample ($\mathscr{R}$, thus the size of set $\mathscr{D}_V$): $DP: = \mathscr{R}_+ / \mathscr{R}$. The idea here is that a useless model would have a low value of *DP*, similar to what would be obtained from setting a ranking of victims purely randomly. Such a model would not be useful at all for the validation sample, no matter how good it could have been for the learning sample.
- *Efficacy Rate*, *E*, which is defined as the ratio between the size of the reserved cases which improve their scores with respect to the initial instance of knowledge ($\mathscr{R}_+$), and the total number of cases which improve their scores ($\mathscr{N}_+$): $E = \mathscr{R}_+ / \mathscr{N}_+$. Hence, heuristically, E measures how *informative* the model, which is selected from the training sample, becomes for the validation sample.

Therefore, a good result achieves high values of both the *Discriminating Power* and *Efficacy Rate*. A cross-validation *realization* is defined as a particular division of sample $\mathscr{D}$ into two sub-samples: a so-called *learning sample* (denoted by $\mathscr{D}_L$) and an evaluating sample (denoted by $\mathscr{D}_V$). Several independent realizations (typically 50) are implemented for each partition of the non-genetic variables. Average of *Discriminating Power* and *Efficacy Rate*, $<DP>$ and $<E>$, are calculated over all the realizations for each partition of the set of non-genetic variables.

Once the best set and partition of non-genetic variables to define cells has been chosen, the methodology is implemented to obtain P_j^i of Eq. (2) using the complete sample $\mathscr{D}$ as a learning sample, since it is desirable to take the maximum advantage of the information provided by all the already-solved cases of the event under consideration.

Figure 1 shows average values of *Discriminating Power* and *Efficacy Rate* for different divisions of the space of non genetic variables into cells, obtained for the Fátima event, averaging 50 independent cross-validation realizations, in each of which 25% of the cases were randomly selected as the validation sample $\mathscr{D}_V$. GeoTemporal partition of the space (black circles) shows the best results, both, in terms of Discriminating Power and Efficacy Rate. Different dots (black circles) correspond to different temporal windows to define cells, from 1 day to 90 days (in steps of 3 days, although in the figure only some of them are labeled). By considering only Temporal variables (empty blue circles) the Discriminating Power is maximum, but Efficacy Rate does not exceed a threshold value around 0.018. TimePolitical partition of the space (violet stars) shows similar results to the GeoTemporal one, although always a little worse. However, at the moment of choosing the best temporal parameter, there are results

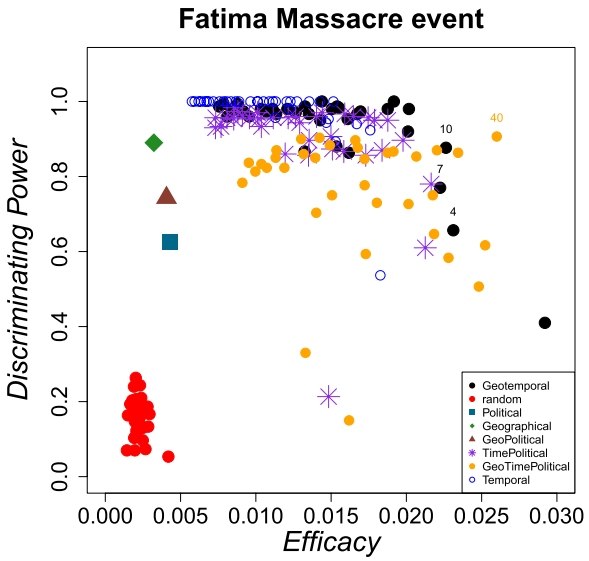

Figure 1. Discriminating Power DP vs Efficacy Rate E by considering different partitions of the space of variables for the Fátima event. Red dots represent the results obtained when the scores for individuals are randomly assigned to the sample. Black circles represent GeoTemporal cells; violet stars, TimePolitical cells; blue circles, Temporal cells; brown triangle, GeoPolitical cells; cyan square, Political cells; green triangle, Geographical cells, and orange circles, GeoTimePolitical cells. In all the partitions involving time variable, each symbol is associated with one temporal window to define the temporal length of the cell (from $T = 1$ day to 90 days), in steps of 3 days.

of GeoTemporal partition that are clearly better than the TimePolitical ones. The green diamond shows the results of using only Geographical variable; the blue square, only Political variable; and the brown triangle (which is approximately in the middle of Political and Geographical ones), using a partition into GeoPolitical cells. The Geographical partition achieves acceptable values of Discriminating Power, although very bad ones in terms of Efficacy Rate. Finally, orange circles represent results by considering the partition which combines the three types of variables, defining GeoTimePolitical cells. Clearly, by considering the three variables, the results are worse than considering only two (PoliticalTime, GeoPolitical and GeoTemporal ones). Red dots represent results when scores for individuals are randomly assigned; they make a well-separated cluster from the rest of the points. The fact that the GeoTemporal and TimePolitical variables lead to results much better than randomly assigned scores shows that there is a knowledge within identified cases of the Fátima event, and then it is possible to take advantage of this knowledge by learning about these non-genetic variables.

From results of Fig. 1, a temporal window of $T = 10$ days using GeoTemporal variables was chosen to define the temporal length of Fátima event´s cells. Then, the methodology was implemented using the complete sample $\mathscr{D}$ as a learning sample, applying the expression (2) to obtain P_j^i values of probability scores for every individual of the set of possible victims V for the selected partition of non-genetic variables (which defines the N_i and I_i values for all i cells, as the example of Table (1)). An example of ranking is shown in Fig. 2. In this Figure, the set of victims is represented on the x-axis; the chosen order in which individuals are represented is that of decreasing ranking scores. The continuous line (red) represents the value for the probability of corresponding with certain UHR from Fátima event at the initial instance of knowledge, before learning about the already-solved cases. Black points represent the probability for every victim in the updated instance of knowledge. In this piecewise-constant distribution, there are no possible victims with zero probability of corresponding with UHR from the Fátima event. This is a consequence of working within a Bayesian framework. The experience learned from the already-solved cases will not lead to establishing possible victims (with non-null probability) and not possible victims (with null probability), but will organize the possible ones hierarchically in terms of probability. Thus, all the possible victims remain in the complete set V after updating the probabilities of the cells, until another type of assumption is made by which some of them are not possible anymore (as in the case of the sex of UHR is known).

Figure 3 shows the results for four mass events of the same context together: Fátima, San Martín, Avellaneda and a Flight event (which was one of the mechanisms applied in Argentina by some IDCs to get rid of people once their death had been decided). In the Figure, each point represents the results of Discriminating Power vs. Efficacy for the selected best partition of variables (GeoTemporal in all cases although with different values of temporal windows to define cells). Fátima event shows the best results (possibly as a consequence of the event characteristics) not only in terms of Discriminating Power and Efficacy Rate, but also in terms of small variance values. However, the results of Fátima overlap with the results of Avellaneda and Flight events in terms of Discriminating Power and Efficacy. Only results of San Martín event have less Efficacy Rate than the rest of the events, but in terms of Discriminating Power, it overlaps with the others too. Another important fact is that the GeoTemporal partition of the space of non-genetic variables is always the best way to detect patterns, although the best temporal windows (the temporal length of the cells) have different values for each event. Values are similar (from 10 to 20 days) possibly because the same phenomenon (the dictatorship in Argentina) underlies all the events.

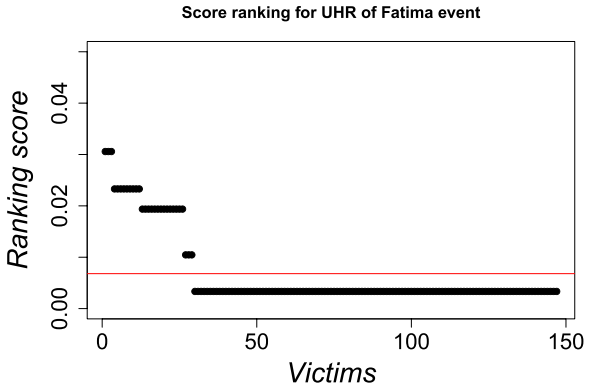

Figure 2. Example of a ranking score for UHR of a woman of (26,40) years old of Fátima event. Victims are represented on x-axes (as a sorted victim index), in decreasing order of the ranking scores. The continuous line (red) represents the value of the probability to correspond with certain UHR from Fátima at the initial instance of knowledge, before learning about the already-solved cases.

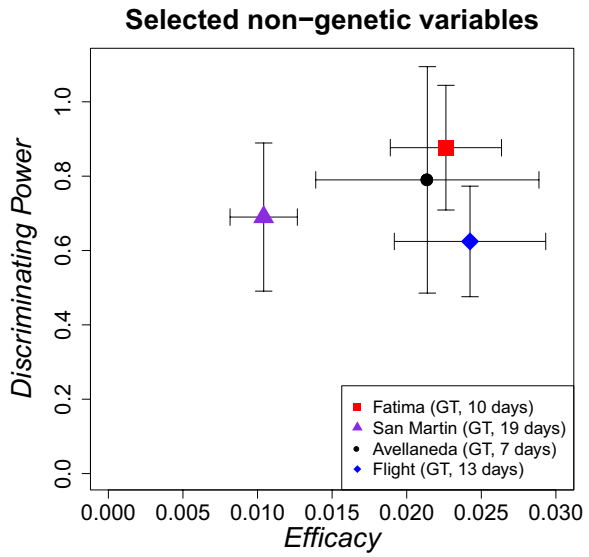

Figure 3. *Discriminating Power DP* vs *Efficacy Rate E* for the selected set of non-genetic variables for the different events under study (Fátima, San Martín, Avellaneda, and Flight events). In all the cases, the best partition is GeoTemporal cells (GT), although with a different temporal parameter to define the length of the cell, depending on the event.

Discussion

For the UHR of a woman from the Avellaneda event, the methodology proposed the fourth position of the score rankings of victims for an individual for whom the EAAF only had a biological reference sample of a distant relative. After these results, the family was contacted to increase the number of reference samples from close relatives. Nowadays she is identified. This reflects the objective of this work: to build a tool that contributes to the work of forensic anthropologists regarding background information, taking advantage of the already-solved cases of a particular event.

The key of this work is very simple: turning to the already-solved cases of a mass event is essential to contribute to the knowledge regarding that event, knowledge that could be used in new searches of the same event, prioritizing the victims for certain UHR, and then, prioritizing the efforts to obtain new ante-mortem data and families' blood samples within the identification process. Prioritization is essential in any investigation of a massive number of victims such as those involving crimes against humanity, conflicts, disaster victim identifications, among others.

Methods

To understand how to update prior probabilities $P\left(S \in \mathcal{N}_j\right)$ into posterior probabilities $P\left(S \in \mathcal{N}_j \middle| \mathcal{D}\right)$, it may be useful to think abstractly about a mathematically equivalent problem, which consists of throwing a die of m faces. It is not possible to assure before collecting any data whether the die is fair (i.e., equal probabilities for every face)

or loaded. The outcomes of a certain number of throws form the data $\mathscr{D}$. Then, the probabilities of each face will be inferred after the observation of a set of outcomes $\mathscr{D}$. In other words, the inference will capture how much unbalanced the die is according to experience. In our case, each face of this die represents each one of the cells of the problem under study. Following this analogy, $P\left(S \in \mathscr{N}_j\right)$ is equivalent to the probability of obtaining face j of the die in the next throw before observing the data $\mathscr{D}$ and $P\left(S \in \mathscr{N}_j|\mathscr{D}\right)$ is the probability of obtaining face j of the die in the next throw after learning about the data $\mathscr{D}$.

In the initial instance of knowledge, there is total uncertainty about which victim is associated with remains S, a fact which is consistent with assigning the same chance to each individual of V. As a consequence, $P\left(S \in \mathscr{N}_j\right) = N_j/N$. The sum of $P\left(S \in \mathscr{N}_j\right)$ over all the cells j is 1, following the assumptions (c) and (d), which means that with certainty UHR S belongs to one of the possible victims of set V. To update $P\left(S \in \mathscr{N}_j|\mathscr{D}\right)$ after learning about the identified cases, it is necessary to go back to the problem of the die, but imagining the experiment of throwing the m faces die n times, being n the total number of already-solved cases in the problem. If face i has a probability θ_i of resulting the winning face, it is possible to compute the probability of the outcome $(n_1, n_2, \ldots, n_m)$ which means that face 1 results n_1 times, and face 2 results n_2 times and so on until the face m, which results n_m times. The probability of obtaining the outcome $(n_1, n_2, \ldots, n_m)$ is a well-known probability distribution called the Multinomial Distribution[19]:

$$P(n_1, n_2, \ldots, n_m|\theta_1, \theta_2, \ldots, \theta_m) = \frac{n!}{n_1!n_2!\ \ldots n_m!}\theta_1^{n_1}\theta_2^{n_2}\ldots\theta_m^{n_m}. \tag{3}$$

It is known that the mean value of the outcome n_1 is $E(n_1) = n\theta_1$, and in general $E(n_i) = n\theta_i$ for the face i. The multinomial distribution can be used to obtain the probability of a particular realization of an experiment in which it is possible to assume the probability of each outcome, or the existence of a model to generate the n outputs for which outcome 1 has probability θ_1, outcome 2, θ_2, and so on up to outcome m, which has probability θ_m. The problem under study is different empirically since there is (obviously) not a die involved, but human beings. Also, it is worth to remember that in Statistical Inference, instead of knowing the true parameters of the probability model, and using them to obtain probabilities of possible different samples, there is a single sample $(n_1, n_2, \ldots, n_m)$ from which to learn the value of the parameters $(\theta_1, \theta_2, \ldots, \theta_j)$. That single sample is the observed data $\mathscr{D}$ from the already-solved cases. In the context of this formalization of the problem, the data $\mathscr{D}$ of the already-solved cases can be thought as a particular realization of the n throws of the die, which results in I_1 identified cases coming from cell 1, I_2 from cell 2, and so on up to I_m identified cases from cell m. Thus, the data $\mathscr{D}$ became one possible result of the experiment of throwing the die, that is $\mathscr{D} = \{I_1, I_2, \ldots I_m\}$, where the sum of all the elements of $\mathscr{D}$ are n_s, which is the total number of identified cases: $\sum_{j=1}^{m} I_j = n_s$.

It is worthwhile to note that thinking about the die as an abstract mathematical problem and the multinomial distribution that arises therein is actually an approximation. That is because a multinomial model for the data is akin to assuming sampling with replacement. In contrast, UHR recognition involves sampling without replacement, which would entail a multivariate hypergeometric distribution. Mathematically, opting for that approximation has the advantage of having a conjugate prior, which greatly simplifies calculations and makes available a posterior distribution and its first two moments as point estimators in closed-form. The alternative, i.e., the multivariate hypergeometric distribution, does not belong to the exponential family of distributions and thus it has no conjugate prior. Bayesian calculations for that model would entail numerical approximations for the arising integrals, or implementation of the Gibbs sampler, which would, in turn, require special computing code for the algorithms and convergence diagnostics. We aim our manuscript at Forensic researchers hoping they will implement the proposed methodology as a tool to prioritize cases and guide the searches. That simplicity in computational terms is the main reason why we adopt the multinomial approximation in this manuscript.

Bayesian inference. Using the Bayes Theorem to update the probabilities in the ideal problem of the m -faced die, posterior probabilities of the parameters θ_i can be written in terms of prior probabilities:

$$f(\theta_1, \theta_2\ldots\theta_m|\mathscr{D}) = \frac{p(\mathscr{D}|\theta_1, \theta_2\ldots\theta_m)\ \cdot\ f(\theta_1, \theta_2\ldots\theta_m)}{P(\mathscr{D})}, \tag{4}$$

where

- $f(\theta_1, \theta_2\ldots\theta_m|\mathscr{D})$ is the posterior density of the parameters $(\theta_1, \theta_2\ldots\theta_m)$,
- $p(\mathscr{D}|\theta_1, \theta_2\ldots\theta_m)$ is the likelihood, which was expressed as a multinomial distribution in this problem,
- $f(\theta_1, \theta_2\ldots\theta_m)$ is the prior density of the parameters $(\theta_1, \theta_2\ldots\theta_m)$,
- and the denominator $P(\mathscr{D}) = \int\cdots\int p(\mathscr{D}|\theta_1, \theta_2\ldots\theta_m)\ \cdot f(\theta_1, \theta_2\ldots\theta_m)\ d\theta_1\ldots d\theta_m$ is the integral (across the prior) of the product of the likelihood and the prior.

A Dirichlet distribution is considered as a prior density of the cell parameters $f(\theta_1, \theta_2\ldots\theta_m)$. By writing a prior distribution of the parameters, it is assumed that the parameters are not fixed values but random variables that have a certain probability distribution, which is expressed as a function of certain hyperparameters $\alpha_1, \alpha_2, \ldots, \alpha_m$. It is given in Eq. (5):

$$f(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma\left(\alpha_1 + \alpha_2 + \ldots + \alpha_j + \ldots\right)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_j)\ldots}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\ldots\theta_m^{\alpha_m-1}, \tag{5}$$

where each $\theta_i \geq 0$ and $\theta_1 + \ldots + \theta_m = 1$.

There are several reasons for which a Dirichlet distribution is used as a prior. The first one is that this distribution adapts to both informative and non-informative prior distributions since it allows the calibration of the hyperparameters for both situations. *Informative* means that not all the possible outcomes are equally likely to occur (as in our case). The second reason is that the Dirichlet is the conjugate prior for the Multinomial distribution, which means that the following property is met: if the likelihood of the parameters is a Multinomial distribution and the prior density of the parameters is a Dirichlet one, then the Posterior density of the parameters $f(\theta_1, \theta_2\ldots\theta_m|\mathscr{D})$ is also a Dirichlet distribution but with updated values of the parameters, which are going to be a known function of the hyperparameters and the data $\mathscr{D}$.

By having a distribution of probability associated, the parameters θ have an associated uncertainty. The following expressions are met for the expected values and variance of the parameters θ_j ($E\left(\theta_j\right)$ and $Var\left(\theta_j\right)$ respectively) as a function of the hyperparameters α_j for all j value:

$$E\left(\theta_j\right) = \frac{\alpha_j}{\alpha_0} \quad Var\left(\theta_j\right) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}, \tag{6}$$

where $\alpha_0 = \alpha_1 + \alpha_2 + \ldots + \alpha_m$. Those properties render a very amenable interpretation for prior elicitation (*i.e.* translating expert knowledge into concrete values)[16]. GeoTemporal cell probabilities at the initial instance of knowledge of the problem have to be consistent with what is known before the already-solved cases. This means that all the possible victims are equally probable, and therefore those cells with more victims will be more likely. As a consequence, before the data from the identifications, the prior probabilities of the cells are proportional to the number of victims belonging to every GeoTemporal cell. A further condition is that the parameters of the distribution are such that the prior expectation of θ_j, $E\left(\theta_j\right) = N_j/N$ for all $j = 1, 2\ldots, m$. In this way, it is required that all individuals have the same chance of corresponding with UHR S in the instance of the knowledge prior to the data of the already-solved cases. As for the Dirichlet distribution it is known that the expected value of the variable θ_j is α_j/α_0, from Eq. (6), then what must be satisfied are m conditions, one for each j cell:

$$E\left(\theta_j\right) = \frac{\alpha_j}{\alpha_0} = \frac{N_j}{N}. \tag{7}$$

The m conditions of Eq. (7) are not independent, since the sum of α_j over all j from 1 to m is α_0. This fact implies that it is not possible to solve the system equations of m unknown variables and only $m - 1$ independent equations. It is necessary to propose another condition as an extra equation. The coefficient of variation (CV) of a random variable is defined as the quotient between its standard deviation and its expected value. As a conventionally accepted "rule of thumb", random variables with a coefficient of variation greater than 0.5 are considered very heterogeneous, while those with values lower than 0.10 are considered very homogeneous. Since it is expected to specify a fairly vague prior, it is possible to establish the coefficient of variation of the most populated cell to be 1, as a criterion to establish the needed extra condition, *i.e.*,

$$\mathrm{CV}\ (\theta_k) = \frac{\sqrt{Var(\theta_k)}}{E(\theta_k)} = 1, \tag{8}$$

where cell k is that cell for which N_k is maximum over all the N_j cell values. The fact that this condition is met for the most populated cell ensures that the coefficient of variation will be more than 1 for the rest of the cells. By squaring the expression of Eq. (8) and replacing it in the second one of Eq. (6), the following extra condition for the hyperparameters is obtained:

$$\frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} = \left(\frac{\alpha_k}{\alpha_0}\right)^2 \tag{9}$$

After solving the system of equations given by the Eqs. (7) and (9), an expression for α_0 as a function of the data is obtained as: $\alpha_0 = \frac{N}{N_k} - 2$. By replacing this result in Eq. (7), an expression for every hyperparameter α_j corresponding to cell j is obtained:

$$\alpha_j = \frac{N_j}{N}\left(\frac{N}{N_k} - 2\right) \quad \forall \quad j \tag{10}$$

It is important to note that if $N_j = 0$ for cell j, then $\alpha_j = 0$, which implies that the expected value is zero ($E\left(\theta_j\right) = 0$) for this cell. In other words, if there are no cases of kidnapped individuals within the place and dates corresponding with the GeoTemporal cell j, then the probability of remains S to be associated to somebody from this cell is null, which makes sense because there are no people with that combination of values of the variables.

Using the property of conjugate distributions for the Dirichlet (prior distribution) and Multinomial (likelihood), the posterior distribution results in a Dirichlet distribution with updated parameters:

$$f(\theta, \theta_2,.., \theta_m) = \frac{\Gamma\left(\alpha_1' + \alpha_2' + \ldots + \alpha_j' + \ldots\right)}{\Gamma(\alpha_1')\Gamma(\alpha_2')\ldots\Gamma\left(\alpha_j'\right)\ldots}\theta_1^{\alpha_1'-1}\theta_2^{\alpha_2'-1}\ldots\theta_m^{\alpha_m'-1} \tag{11}$$

where values of $\alpha_1', \alpha_2', \ldots, \alpha_m'$ are functions of the hyperparameters α and the data $\mathscr{D}$, as: $\alpha_j' = \alpha_j + I_j$, in general for cell j. This means that the expected values of the posterior parameters $\theta_1, \theta_2, \ldots, \theta_m$ are:

$$E^{post}\left(\theta_j\right) = \frac{\alpha_j'}{\alpha_0'} = \frac{\alpha_j + I_j}{\alpha_0 + n_s}$$

These are the Bayesian estimators of the unknown probabilities which use both the knowledge given by the prior distribution (equiprobability of all the eligible victims) and the observed data (the set of already-solved cases) for a particular cell partition of non-genetic variables. This means that in the problem of the die the probability that the outcome is face j (the cell j in the problem at hand) will depend on the prior expected value modified by the data $\mathscr{D}$ from the already-solved cases (both of the total already-solved cases identified n_s and of the total of already-solved cases that fell into that cell, I_j). These results imply that the probability that remains S belong to some victim of cell j, which is $P\left(S \in \mathscr{N}_j \middle| \mathscr{D}\right)$ (called θ_j^{post} for the sake of simplicity), is:

$$P\left(S \in \mathscr{N}_j \middle| \mathscr{D}\right) = \frac{\alpha_j + I_j}{\alpha_0 + n_s} = \frac{\frac{N_j}{N}\left(\frac{N}{N_k} - 2\right) + I_j}{\frac{N}{N_k} - 2 + n_s} \equiv \hat{\theta}_j^{\text{post}}$$

This work was accepted and presented in the Congress of the American Academy of Forensic Sciences[20]. Data are not available but all the scripts for implementing the methodology in R are available in https://github.com/inescaridi/PriorID as the project *priorID*. An open, free, multi-platform and standalone interface for users to implement this methodology in diverse problems and incorporate feedback from forensic researchers is now under construction.

Received: 1 August 2019; Accepted: 18 December 2019;
Published online: 19 March 2020

References

1. Penchaszadeh, V. Use of dna identification in human rights, work to reunite families in latin america, in els, john wiley & sons, ltd (2001).
2. Cordner, S. & McKelvie, H. Developing standards in international forensic work to identify missing persons. *Int. Rev. Red Cross* **84**(848), 867–884, https://doi.org/10.1186/2041-2223-2-15 (2002).
3. Donkervoort, S., Dolan, S. M., Beckwith, M., Northrup, T. P. & Sozer, A. Enhancing accurate data collection in mass fatality kinship identifications: lessons learned from hurricane katrina. *Forensic Sci. Int. Genet.* **2**, 354–362 (2008).
4. Baeta, M. *et al.* Digging up the recent spanish memory: genetic identification of human remains from mass graves of the spanish civil war and posterior dictatorship. *Forensic Sci. Int. Genet.* **19**, 272–279 (2015).
5. Penchaszadeh, V. B. Forced disappearance and suppression of identity in children of argentina: Experiences after genetic identification. en: S.gibbon, r. ventura santos, m. sans (eds). racial identities, genetic ancestry and health in latin america. london: Palgrave mcmillan (2011).
6. Dolan, S. M. *et al.* The emerging role of genetics professionals in forensic kinship dna identification after a mass fatality: lessons learned from hurricane katrina volunteers. *Genet. Medicine* **11**(6), 414–7, https://doi.org/10.1097/GIM.0b013e3181a16ccc (2009).
7. Evett, I. W. & Weir, B. S. Interpreting dna evidence, statistical genetics for forensic scientists, the forensic science service. united kingdom (1998).
8. Budowle, B., Ge, J., Chakraborty, R. & Gill-King, H. Use of prior odds for missing persons identifications. *Investig. Genet.* **2**, 15, https://doi.org/10.1186/2041-2223-2-15 (2011).
9. Ge, J., Budowle, B. & Chakraborty, R. Choosing relatives for dna identification of missing persons. *J. Forensic Sci.* **56**, S23–8, https://doi.org/10.1186/2041-2223-2-15 (2011).
10. Vullo, C. M. *et al.* Ghep-isfg collaborative simulated exercise for dvi/mpi: Lessons learned about large-scale profile database comparisons. *Forensic Sci. Int. Genet.* **21**, 45–53, https://doi.org/10.1186/2041-2223-2-15 (2016).
11. Baker, L. E. & Baker, E. J. Reuniting families: An online database to aid in the identification of undocumented immigrant remains. *J Forensic Sci.* **53**, 50–3, https://doi.org/10.1111/j.1556-4029.2007.00612.x (2008).
12. http://www.eaaf.org/.
13. SaladoPuerto, M. & Tuller, H. Large-scale forensic investigations into the missing: Challenges and considerations. *Forensic Sci. Int.* **279**, 219–228, https://doi.org/10.1016/j.forsciint.2017.08.025 (2017).
14. Caridi, I., Dorso, C. O., Gallo, P. & Somigliana, C. A framework to approach problems of forensic anthropology using complex networks. *Phys. A.* **390**, 1662, https://doi.org/10.1016/j.physa.2010.11.042 (2011).
15. I., C., E., A., C., S. & M., S. P. A new complex investigation model for searching, mapping, and identifying disappeared persons in argentina, proceedings of the american academy of forensic sciences, new orleans, luisiana, usa (2017).
16. O'Hagan, A. *et al.* Uncertain judgements: Eliciting experts' probabilities, statistics in practice, wiley, nueva york, ee.uu. (2006).
17. Faraway, J. J. Does data splitting improve prediction? *Stat. Comput.* **26**, 49–60, https://doi.org/10.1007/s11222-014-9522-9 (2016).
18. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning, data mining, inference and prediction, 2nd edition, springer-verlag new york inc. (2009).
19. Johnson, N. L., Kotz, S. and Balakrishnan, N. Discrete multivariate distributions univariate discrete distributions, set, 3rd edition (1997).
20. Caridi, I., Alvarez, E., Somigliana, C. & SaladoPuerto, M. A new complex investigation model for searching, mapping, and identifying disappeared persons in argentina. new orleans, luisiana, usa,. priorid: Prioritizing victims open, free, multi-platform and standalone interface which implement the methodology is under construction (2017).

Acknowledgements
Authors thank the EAAF Team for compiling and analyzing the data that was used in this study.

Author contributions
I.C., C.S. and M.S.P. conceived the analysis. C.S., M.S.P. and the EAAF Team compiled all the data and information regarding the events. I.C. and E.E.A. implemented the calculations. I.C. processed the data. All authors analysed the results and wrote and reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to I.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.